



# Algorithms 1

We can count each appearance of each 6-100+ nucleotides long subsequence in a genome of any size in minutes and store it in specially designed data structures;

n	4n	0 MM	1 MM	2 MM	3 MM	4+ MM
19	2.75E+11	1.51%	36.00%	57.55%	4.95%	0.00%
20	1.10E+12	0.40%	15.66%	64.79%	19.02%	0.13%
21	4.40E+12	0.10%	5.28%	49.78%	41.76%	3.07%
22	1.76E+13	0.02%	1.55%	26.65%	56.96%	14.82%
23	7.04E+13	0.01%	0.44%	10.54%	49.96%	39.05%
24	2.81E+14	0.00%	0.11%	3.44%	29.25%	67.20%
25	1.13E+15	0.00%	0.03%	1.02%	12.37%	86.58%
26	4.50E+15	0.00%	0.01%	0.29%	4.25%	95.45%
27	1.80E+16	0.00%	0.00%	0.08%	1.33%	98.58%
28	7.21E+16	0.00%	0.00%	0.02%	0.39%	99.59%
29	2.88E+17	0.00%	0.00%	0.01%	0.11%	99.88%
30	1.15E+18	0.00%	0.00%	0.00%	0.03%	99.97%
31	4.61E+18	0.00%	0.00%	0.00%	0.01%	99.99%

19-31-mer  
presence/  
absence statistics  
for the human  
genome



# Algorithms 1

We can count each appearance of each 6-100+ nucleotides long subsequence in genome of any size in minutes and store it in specially designed data structures;

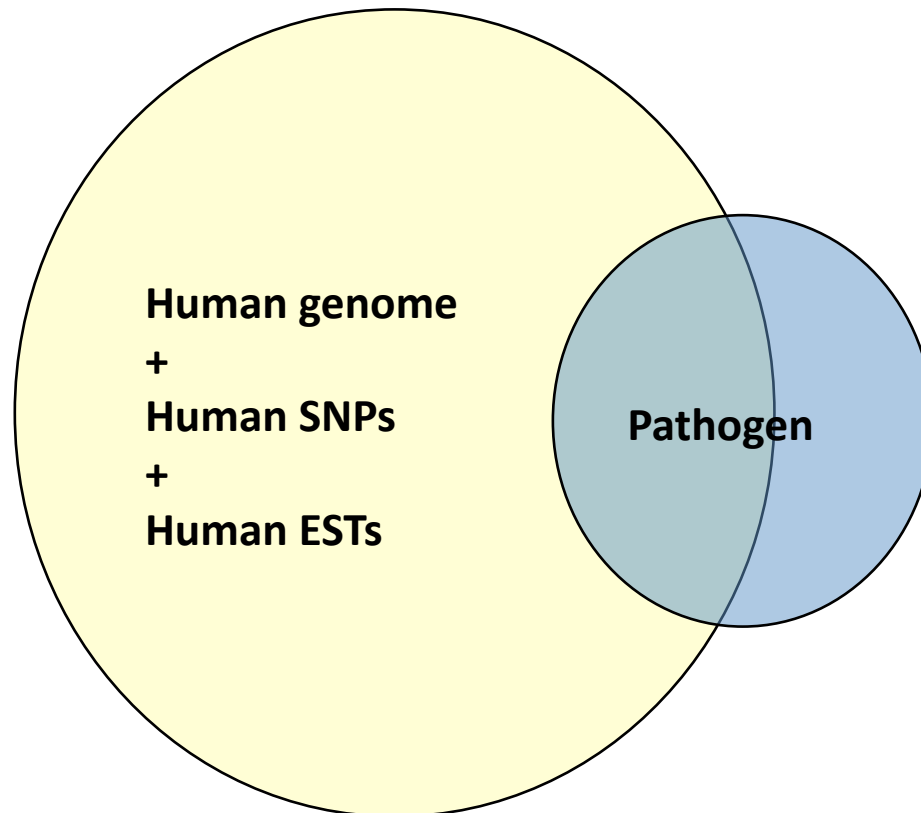
n	0 MM
19	4,126,758,110
20	4,362,754,127
21	4,491,286,106
22	4,574,995,908
23	4,639,348,322
24	4,694,139,197
25	4,743,164,202
26	4,788,126,430
27	4,829,847,940
28	4,868,788,468
29	4,905,330,154
30	4,939,757,090
31	4,972,283,888

Unique  
19-31-mers  
present in the  
human genome



## Algorithms 2

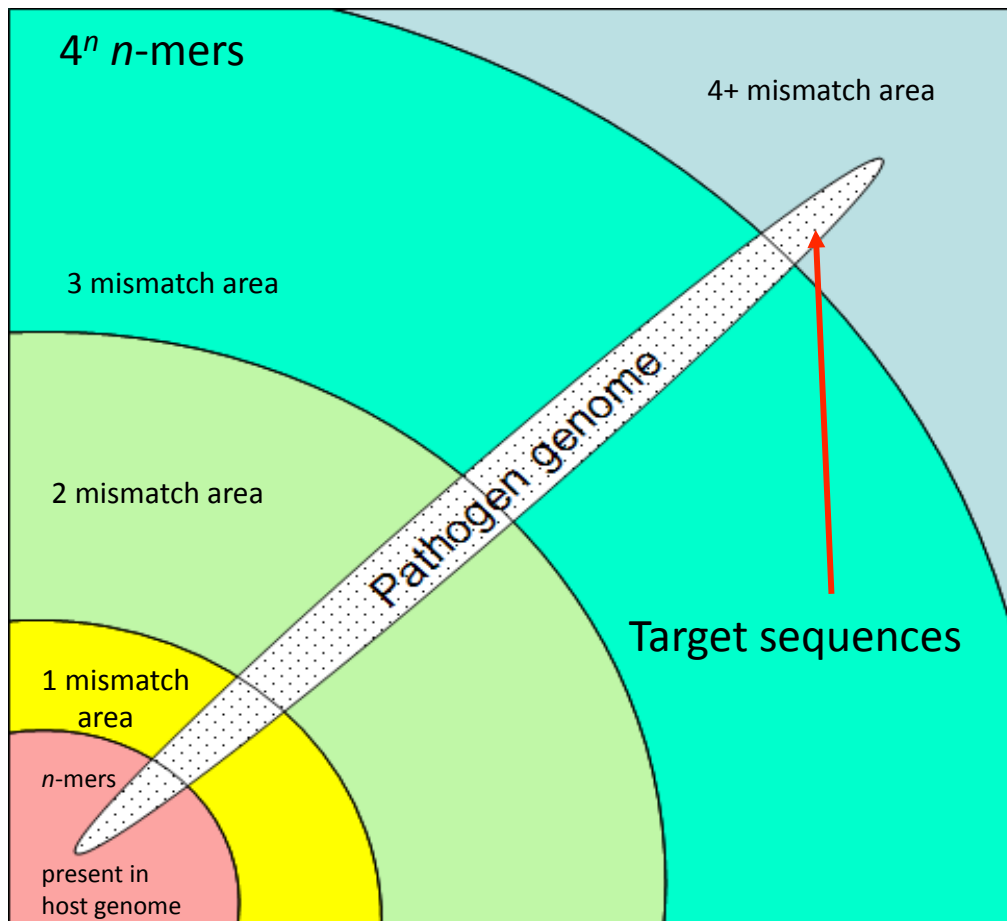
We can do algebra on such data structures in seconds (bacterial genomes) or minutes (human genomes);





# Algorithms 3

We can count each appearance of each subsequence which may appear from each sequence in genome by any combination of 1, 2, 3, and 4+ mismatches;

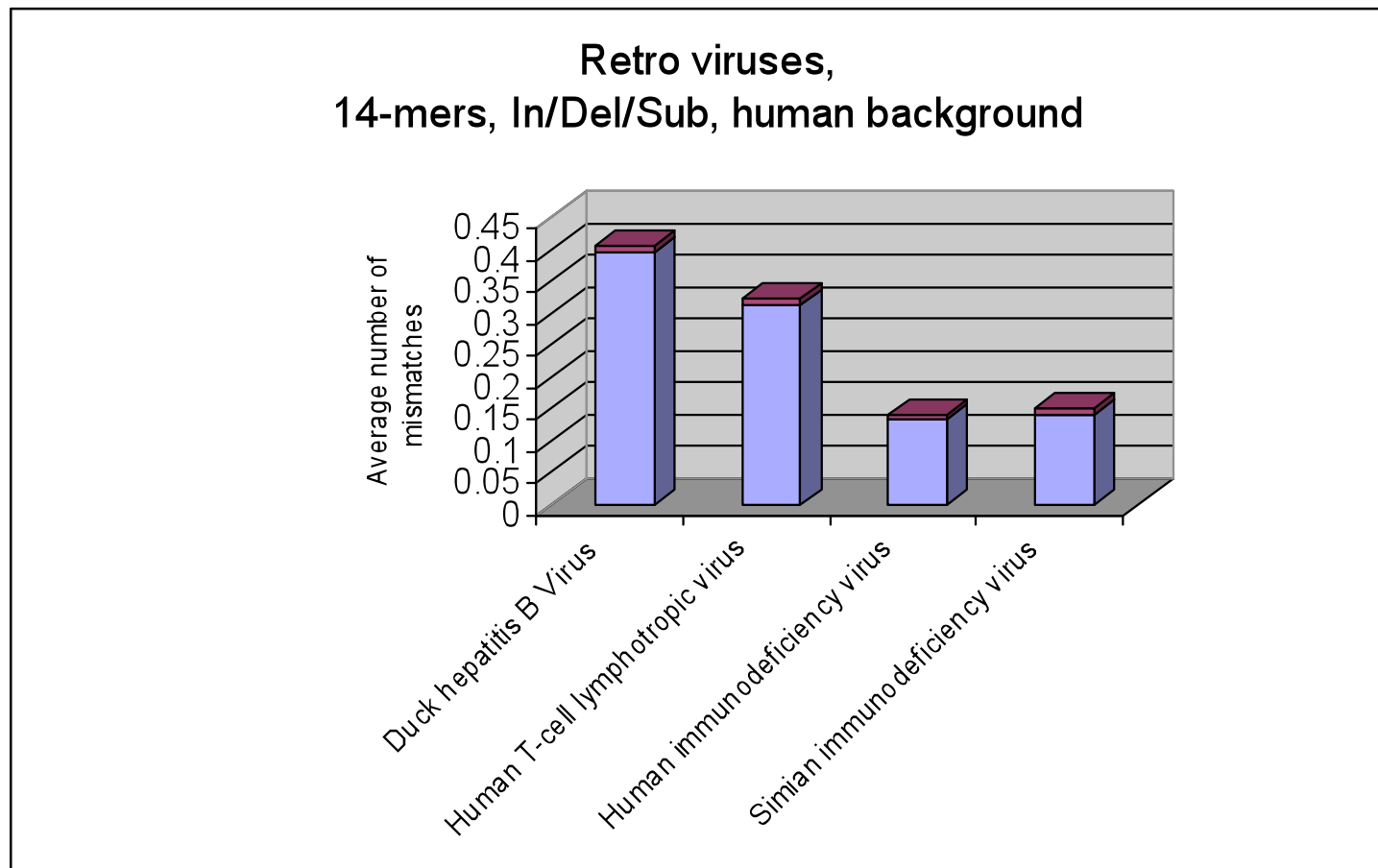


**Including any combination of insertions, deletions, and substitutions on any position**



# Algorithms 4

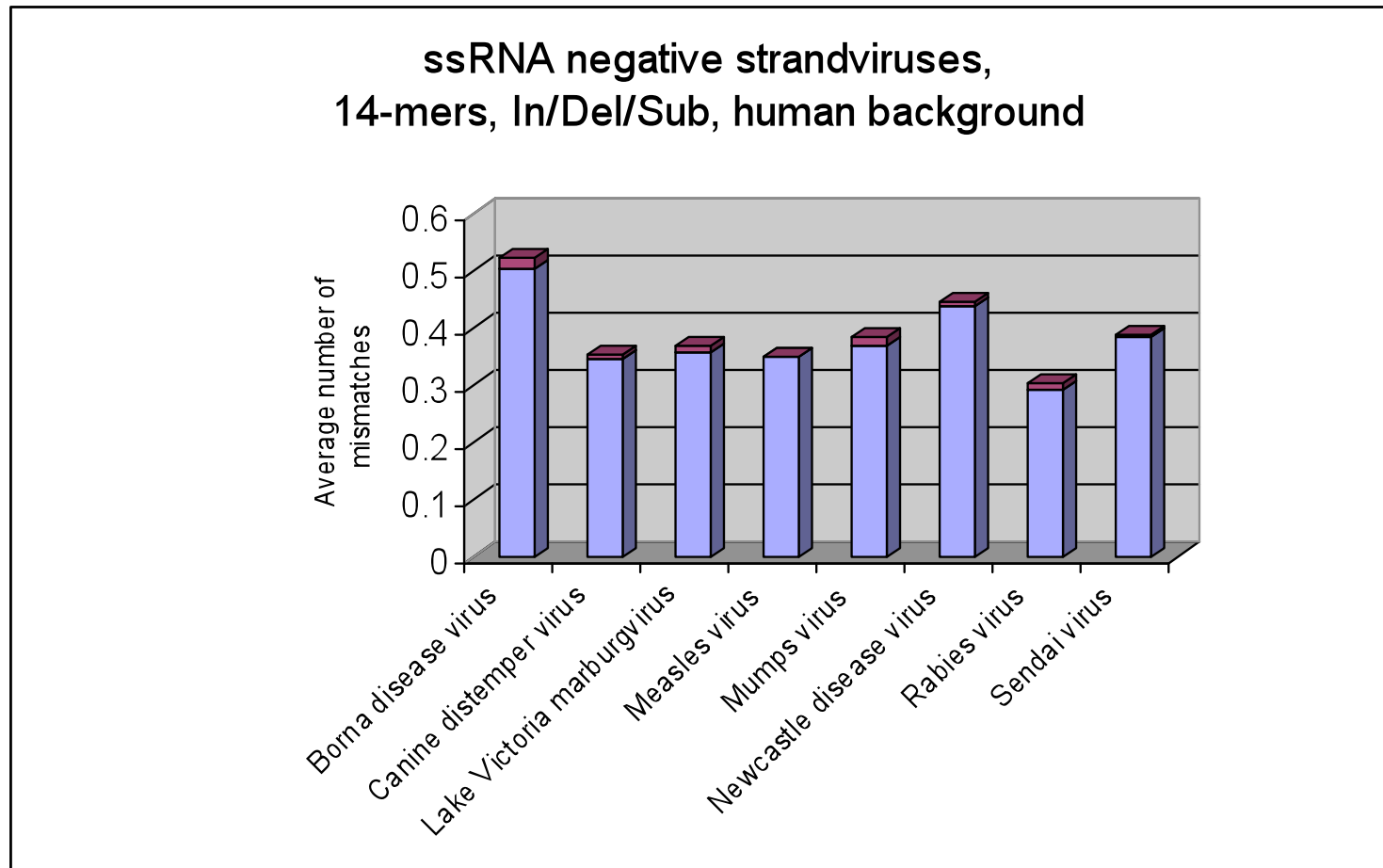
We can identify the average distance of each target genome from the “background” such as human DNA + human SNPs + human ESTs;





# Algorithms 4

We can identify average distance of each target genome from the “background” such as human DNA + human SNPs + human ESTs;

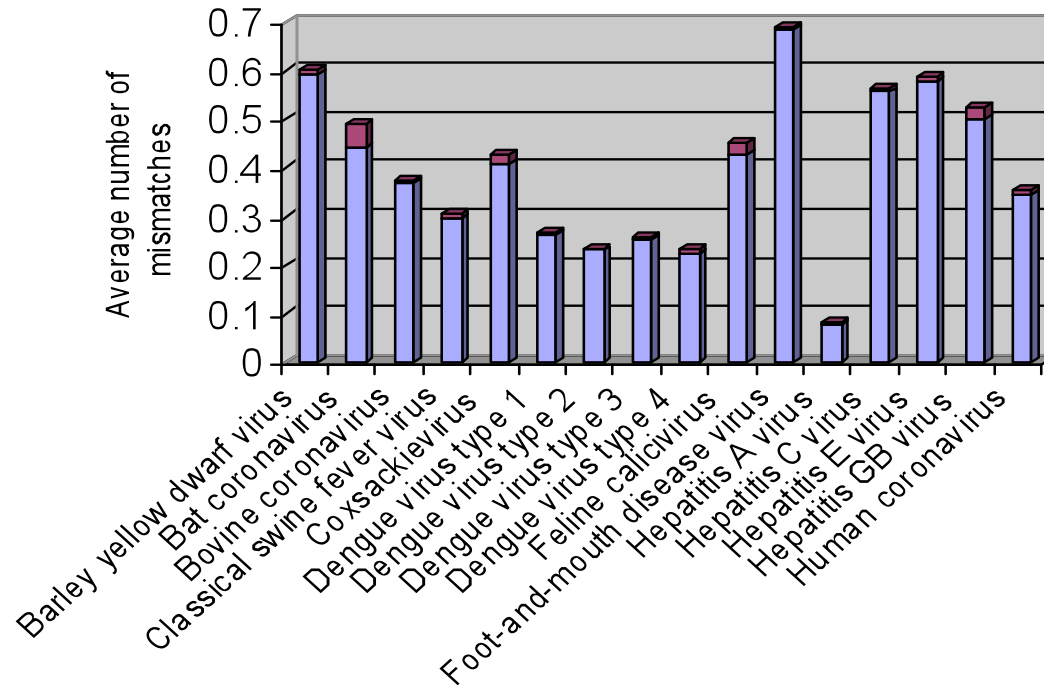




# Algorithms 4

We can identify average distance of each target genome from the “background” such as human DNA + human SNPs + human ESTs;

ssRNA (1),  
14-mers, In/Del/Sub, human background

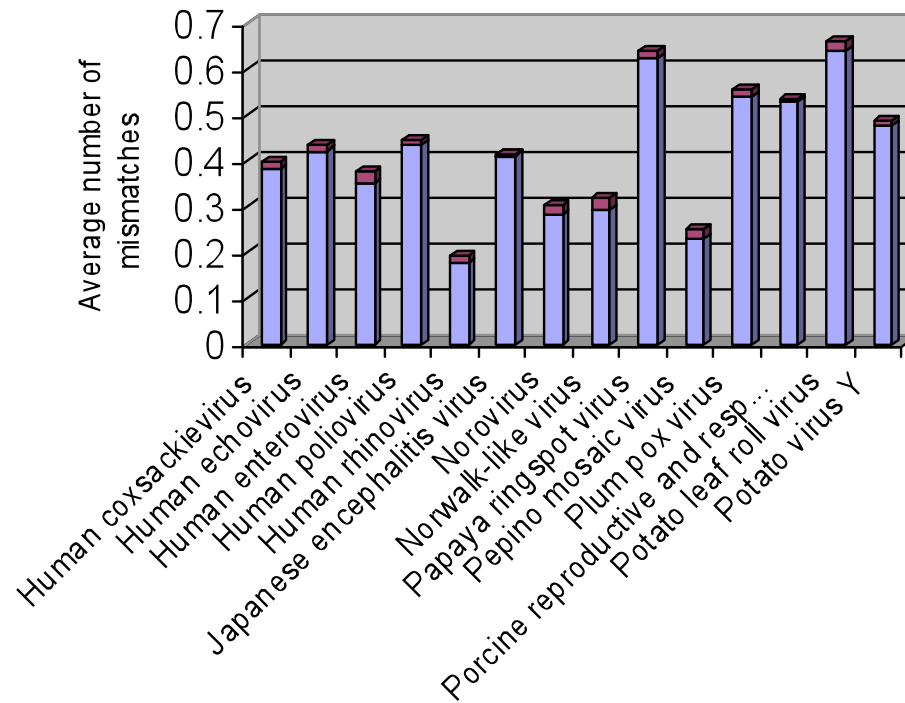




# Algorithms 4

We can identify average distance of each target genome from the “background” such as human DNA + human SNPs + human ESTs;

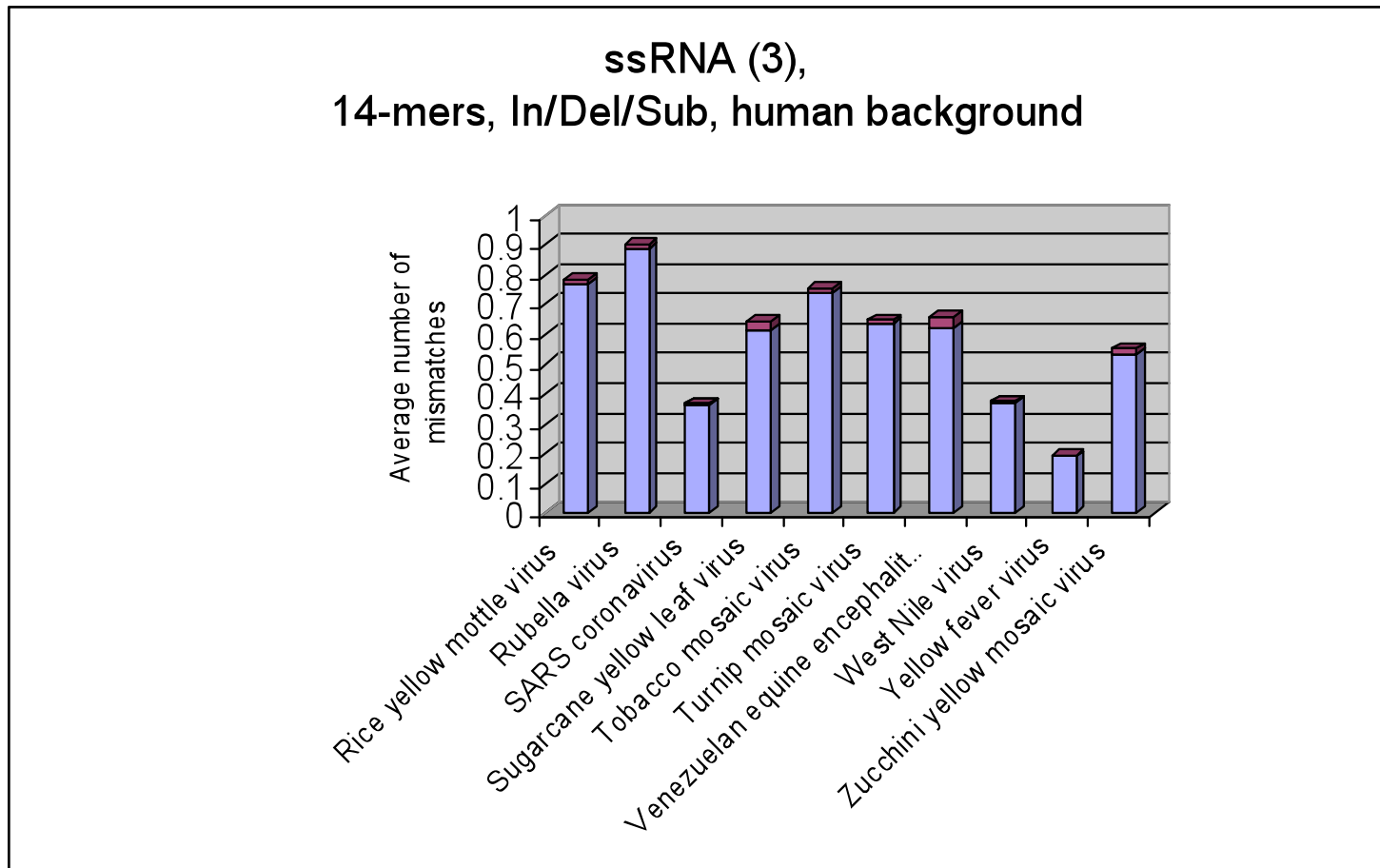
ssRNA (2),  
14-mers, In/Del/Sub, human background





# Algorithms 4

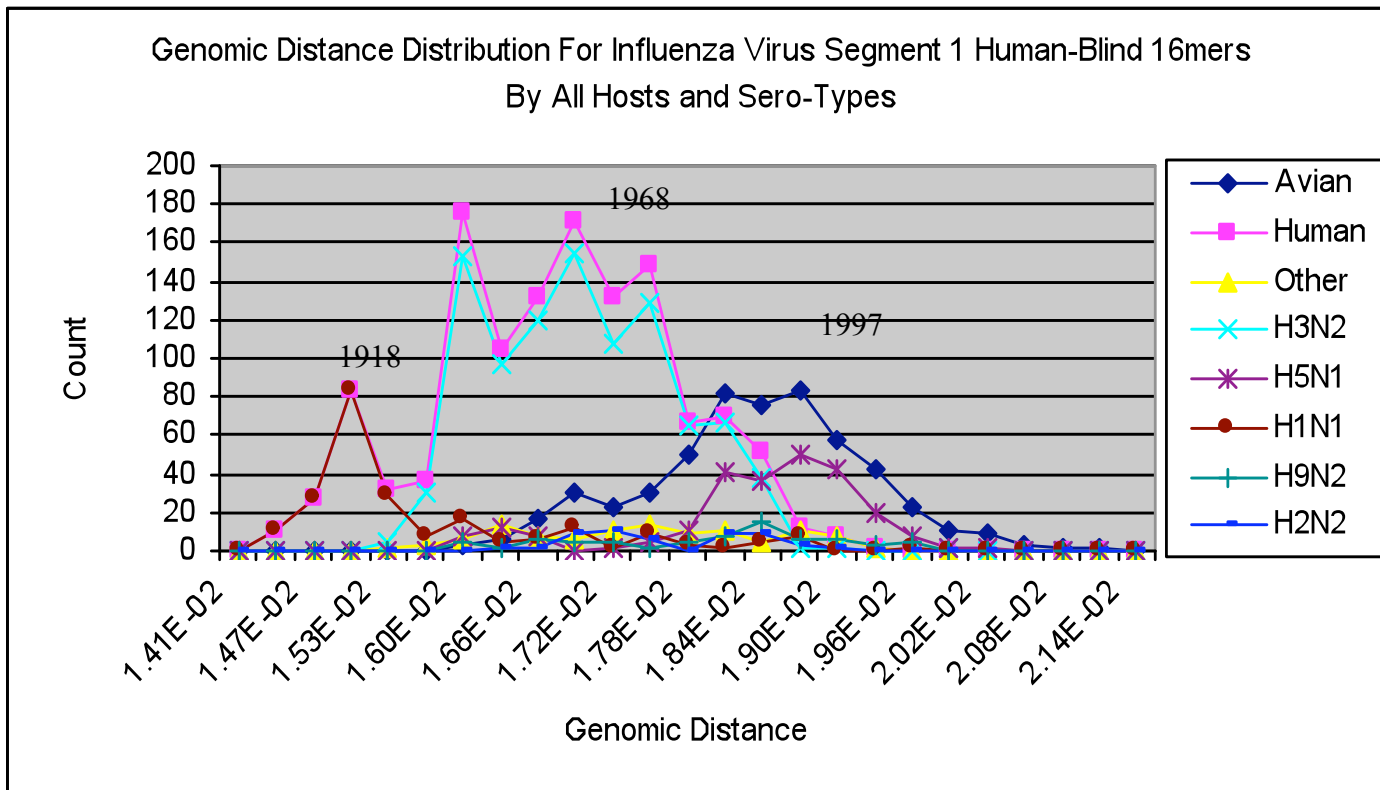
We can identify average distance of each target genome from the “background” such as human DNA + human SNPs + human ESTs;





# Algorithms 5

Identify location and of each potential signature in genome and average distance of each region of the genome from the “background”;



- H1N1: 1918 Spanish flu
- H2N2: 1957 Asian flu
- H3N2: 1968 Hong Kong flu
- H5N1: 1997 avian flu

Influenza A Segment 1 (PB1) Human-Blind 16mers – All Hosts And Serotypes  
Flu: pathogen-host adaptation?



# Algorithms 6

We can identify the minimal combination of background-blind (2+ mismatches away) signatures (probes or primer pairs) needed to identify any set of targets (such as all of the 857 sequenced HIV genomes)

<b>Required coverage</b>	<b>Number of distinguished probes</b>	<b>Number of distinguished primer-pairs</b>
1	6	15
5	27	96
10	68	388
20	206	1690 (622 completely covered strains)



# Algorithms 7

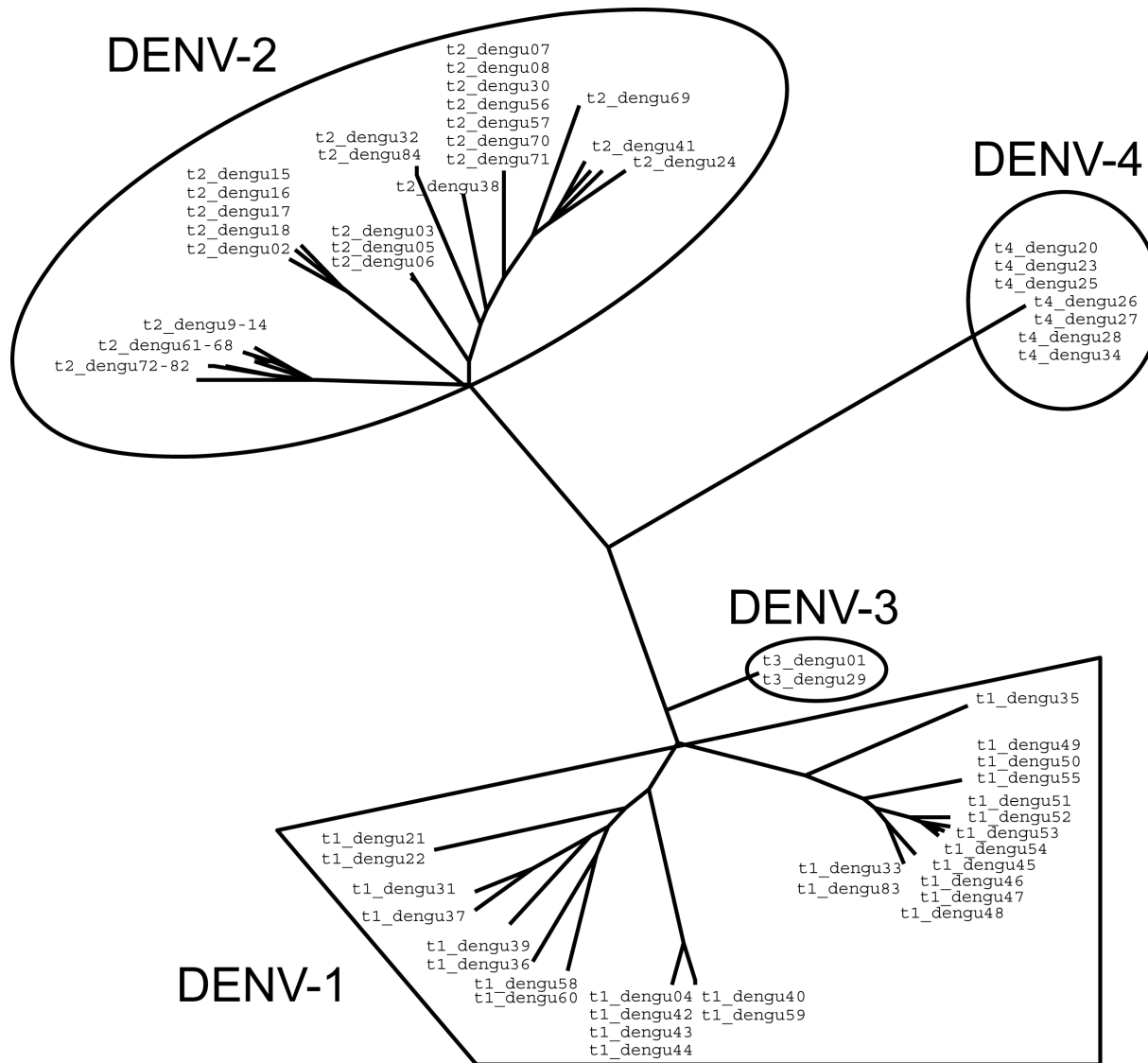
We can identify the minimal combination of background-blind (2+ mismatches away) signatures (probes or primer pairs) needed to identify any set of targets (such as all of the 193 sequenced dengue virus genomes) and to identify any subset (cluster) of targets (such as distinguish between 4 serotypes of dengue virus).

Required coverage	Number of distinguished signatures	Number of primer-pairs require for detection only	Number of primer-pairs require for detection and identification
1	1	5	7
5	5	26	29
10	10	57	50
20	20	127	118

48 serotype 1  
73 serotype 2  
57 serotype 3  
15 serotype 4

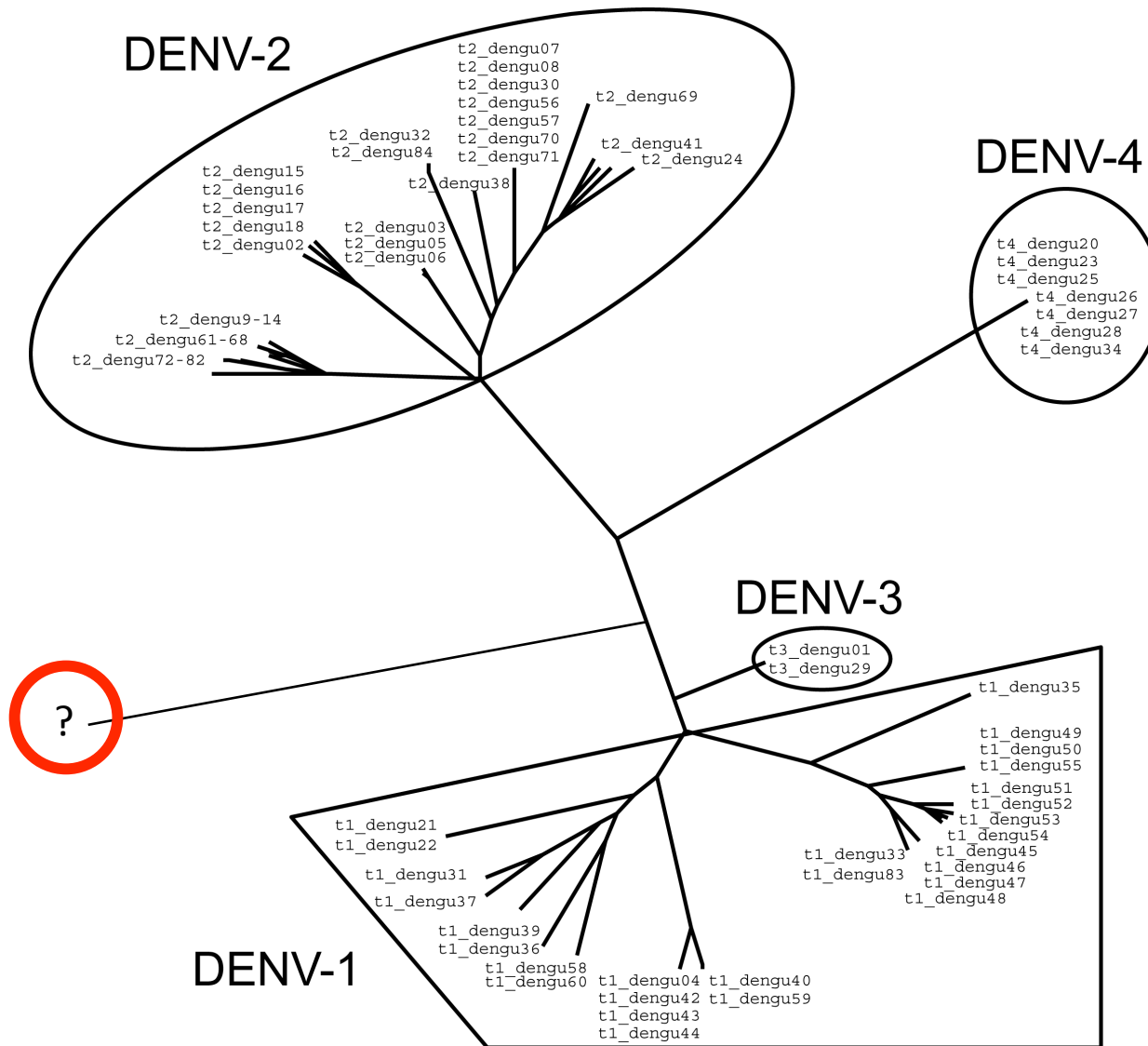


# Example: Host Blind Dengue Virus Microarray





# Example: Host Blind Dengue Virus Microarray



**NEW approach  
to avoid  
False positive?**



# Algorithms Summary – We can:

1. Count each appearance of each 6-100+ nucleotides long subsequence in genome of any size in reasonable time and store it in specially designed data structures.
2. Do algebra on such data structures in seconds (bacterial genomes) or minutes (human genomes);
3. Count each appearance of each subsequence which may appear from each sequence in genome by any combination of 1, 2, 3, and 4+ mismatches, Including any combination of insertions, deletions, and substitutions.
4. Identify average distance of each target genome from the “background” such as human DNA + human SNPs + human ESTs.
5. Identify location of each potential signature in genome and average distance of each region of the genome from the “background”.
6. Identify the minimal combination of background-blind (2+ mismatches away) signatures (probes or primer pairs) needed to identify any set of targets.
7. Identify the minimal combination of background-blind (2+ mismatches away) signatures (probes or primer pairs) needed to identify any subset (cluster) of targets.



Contact

Didier Perez

[didier@eurekagenomics.com](mailto:didier@eurekagenomics.com)

415-269-0666